

## 新的坑

### 一、隐私问题：

统计数据库中的信息在关联之后（如下表），可能会让别人识别出个人信息，从而暴露隐私。

年龄	性别	爱好	工作	个数
XXX	XXX	XXX	XXX	1
XXX	XX	XX	XX	23

### 二、背景、使用环境：

统计数据（医疗数据）管理者在进行展示的时候需要对数据进行处理。

### 三、现有方法：

（优化目标）

- 1、k-anonymity：让每个组中的敏感数据数量大于 k 个。虽然 k 个数据项都存在于一个集合中无法区分，但它们的其他属性相同时失效
- 2、l-diversity：让每个组中的敏感数据都有 l 个以上不同项。有极大相关项辅助预测时失效。
- 3、t-closeness：让每个组中的敏感数据分布和全集分布差别小于 t。分布差距测量方法不能保证有效
- 4、 $\epsilon$ -differential privacy：给数据加噪音。

为了让数据使用性增加，即分组尽可能细致，实现前三个优化目标是 NP 问题。同时在实现过程中，用户可能根据要展现内容的不同可能会有具体的要求（比如一些疾病在不同年龄段有不同程序比例，这种异常需要展示或者是公众常识，则没必要要求这类分布和全体数据近似）。第四个需要对数据项进行修改。以上四个方法都需要先定义参数值。

### 三、我们想做的是：

通过可视化多个属性交集组的数据项数量、敏感数据分布和全部数据分布的区别来让用户知道哪里可能会涉及隐私问题并结合以上理念、用户需求和专家知识来通过交互有效的避免隐私问题。

我们希望根据用户体验，一个一个或一组组的选择属性取交，可视化出集合不断分裂的过程，高亮出可能涉及隐私问题的交集，并提供交互可以让用户将属性重新划分（合并拆分等等）

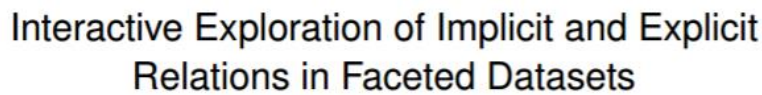
### 四、分析流程：

（暂无理论支持）

观察数据属性→选择一些属性→观察在一定分组条件下各交集组内的敏感数据→调整分组→添加属性→观察→调整…

### 五、有关系统：

Cong Xie, Wen Zhong, and Klaus Mueller, *Senior Member, IEEE*



用平行坐标的线表示一个数据交集组，联合其他可视化展示该组的数据项个数、分布等等。用户可以在平行坐标上交互地更改分组方式，同时可以增减坐标轴，即数据属性，从不同粒度观察交集组。

t-Closeness: Privacy Beyond k-Anonymity and l-Diversity  
t-Closeness through Microaggregation: Strict Privacy with Enhanced Utility Preservation  
Differential Privacy via t-Closeness in Data